Genetic Subtypes of HIV-1

Thomas Leitner

Theoretical Biology and Biophysics, Group T-10, MS K710,
Los Alamos National Laboratory, Los Alamos NM 87545

Current Subtypes

HIV-1's are currently divided into two genetic groups, based on phylogenetic reconstruction using DNA sequences. The majority of these sequences fall into the M (major) group, while a smaller, but growing, number of sequences are classified as O (outlier). The M group has been further subdivided into several subtypes formed by more or less distinct clades in the M group phylogeny. These clades have been given subtype classification names from A to J (Figure 1).

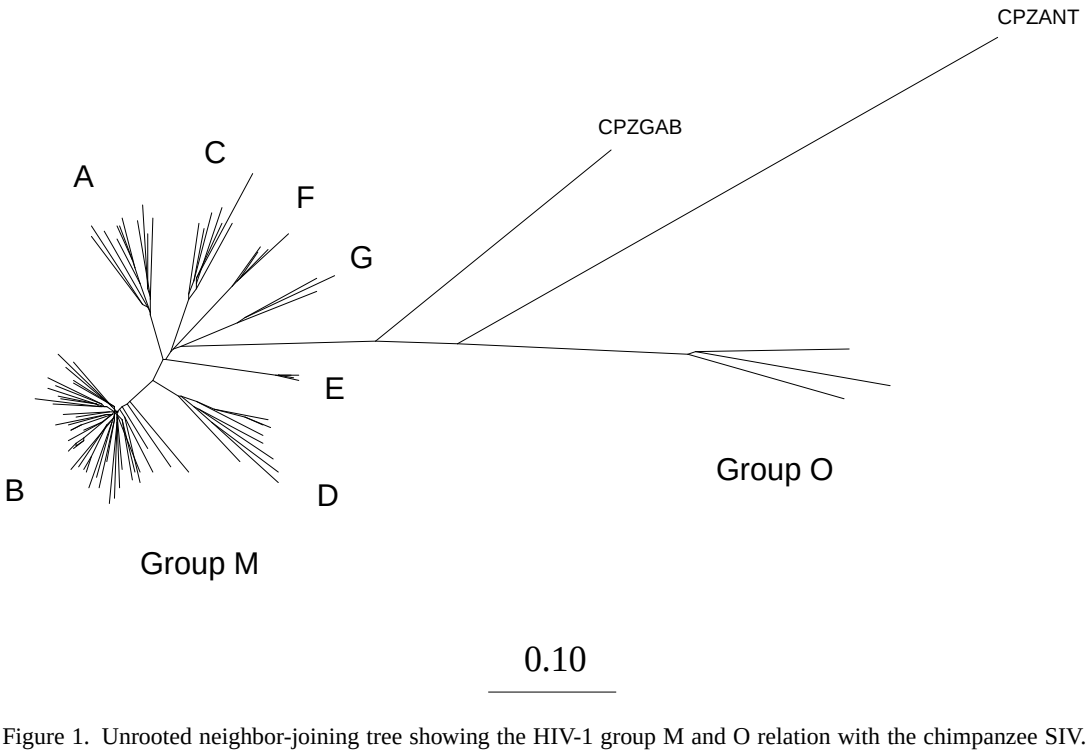

Figure 1. Unrooted neighbor-joining tree showing the HIV-1 group M and O relation with the chimpanzee SIV. The group M subtypes A–G are also indicated at the respective cluster, while for subtypes H and J full length sequences do not yet exist and therefore they are not included in this tree. The tree was based on all nucleotide sequences corresponding to the full length gp120 coding region of *env* in the data base, using a distance matrix calculated with the F84 (κ = 2.0) substitution model [19].

In phylogenetic trees the chimpanzee sequences usually have a branch point located between Group M and O (Figure 1), suggesting distinct origins of the Group M and O viruses. The branching order of the two chimpanzee sequences (CPZGAB and CPZANT) may change when different sequence fragments are investigated. The explanation for this inconsistency may be recombination between chimpanzee sequences.

Definition of a Subtype

The bases for what should be called a genetic subtype were originally proposed by Myers in *Human Retroviruses and AIDS 1994* [47].

i) subtypes are approximately equidistant from one another in *env* (a "star phylogeny");

ii) the *env* phylogenetic tree is for the most part congruent with the *gag* phylogenetic tree;

iii) two or more samples are required to define a sequence subtype.

These are both logical and reasonable practical criteria for the subdivision of sequences based on phylogenetic trees. However, many sequences were found to be unclassified because they clustered differently using different phylogenetic reconstruction methods or in different reference data sets. Furthermore, as discussed below, some sequences were either too short or were actually just one representative of a potential subtype. With recent knowledge and methods of detecting intersubtype recombinants, many of these unclassified sequences have been explained. In fact, many of these problems are typical for chimeric sequences. Salminen et al., therefore, in the 1995 issue of the *Human Retroviruses and AIDS* [55] suggested that the criteria for new subtypes should include:

i) the appearance of at least two epidemiologically unrelated isolates that cluster together and are separated from established genotypes;

ii) the availability of at least 1.5 kilobases of contiguous sequence from each;

iii) the absence of any subsegment that can join established genotypes.

The sum of all these criteria form a reasonable sound basis for establishing a new subtype. However, when attempting to classify a sequence as an established subtype, less strict criteria may be used, as long as this is made clear. For instance, if only the *env* V3 region (typically 300 bp) was used to classify a virus one might propose that the virus is *env* V3 subtype X (where X = A, B, C, etc.). It would be better, of course, if more of the above criteria were fulfilled. Should the sequence fall outside any of the established clusters, it is desired that further sequence be generated and that additional analyses be performed, as described below.

The M Group

Table 1 lists typical genotypes from each of the group M HIV-1 subtypes. The sequences in this table were chosen accordingly to these criteria: (1) The sequence from the sample should exist in *env*, *gag*, and *pol*; (2) The sequence should be a full length gene sequence; (3) The sequences should representthe variation within each subtype; (4) The sequences should not show any signs of recombination in any of the three genes *env*, *gag*, or *pol*; and (5) A maximum of four sequences should represent each subtype. These criteria were chosen in order to facilitate the subtyping effort. For instance, criterion 1 makes it easier to compare results from different genomic fragments, and criterion 2 makes it possible to use the same references even though studies are based on different subfragments. Criteria 3 and 4 allow a researcher to detect odd sequences for further analysis (see below), and finally, criterion 5 lessens the computational effort required. However, these criteria were not possible to follow in every case: in the more uncommon subtypes (G, H, and J) complete genes have not always been sequenced. Subtype I has not been included in the list, because it is only determined in the *env* V3 region [30], and has also preliminary been proposed to be a recombinant [53]. In *pol*, less sequencing has been done; therefore, the criteria were followed less strictly for this gene. Subtype E is also an exception from criterion 1 [23].

Table 1 also gives a brief description of the principal geographical locales of each subtype. All subtypes have been found in Central Africa, while only one or a few subtypes dominate in other parts of the world. However, in some European countries several subtypes have been reported [4, 21, 30, 54], and in one case nearly all subtypes have been found [1].

Subtype A is the most genetically diverged subtype, and in many tree analyses the more rare subtypes G and H are close to subtype A. Furthermore, all subtype E and at least three subtype G viruses have been found to cluster in subtype A in the 3' half of their gp41 coding region, suggesting that they are recombinants [23]. The Thai envelope subtype E sequences all appear to be mosaic, with

Table 1 Proposed Reference Sequences of HIV-1 Genetic Subtypes

Subtype	*env*	*gag*	*pol*	Main geographical area
A	U455	U455	U455	Central Africa
	IBNG	IBNG	IBNG	
	DJ258	DJ258		
	SF1703	VI32		
B	LAI	LAI	LAI	Europe
	JRFL	JRFL	JRFL	North and South America
	OYI	OYI	OYI	Australia
	RF	RF	RF	Asia
C	UG268	UG268		East and South Africa
	ZAM18	ZAM18		
	ZAM20	ZAM20		
	DJ256	DJ259		
D	NDK	NDK	NDK	Central Africa
	JY1	VI203	Z2Z6[1]	
	UG274	UG274	ELI[1]	
	SE365	UG270		
E	TN235			East Asia
	TN239			
	TN242			
	CM240X			
F	BZ163A	BZ162		South America
	BZ126A	VI69		East Europe
	93BR029.2	VI174		
	93BR020.17			
G	LBV217[1]	LBV217[1]		Central Africa
	92UG975.10	VI191		
	92RU131.9	SE6165[2]	SE6165[2]	
H	VI557[2]	VI557		Central Africa
	CA13[2]	VI525		
J	SE702[2]	SE7022[2]		Central Africa
	SE7887[2]	SE7887[2]		(Europe)

[1] These sequences have been found to be partly recombinant in *env* or *gag*.

[2] Full length gene sequences do not yet exist.

the *gag* gene from subtype A and the *env* from subtype E. The parental subtype E virus has yet not been found, but recently an extensive analysis of one such mosaic isolate showed multiple breakpoints between the two parental genotypes [9]. Subtype A is mostly prevalent in Central Africa, but has also been found in other parts of the world, including Europe [1, 4], Russia [41], East Asia [65], and America [7, 43]. It is possible that this subtype represents a contemporary pool of the original cross-species transmission(s) from monkey to human, from which all the other subtypes have been drawn [34].

Subtype B is by far the most studied subtype, mainly because viruses from this subtype represent the vast majority circulating in the western world, but also because many original laboratory isolates are of this type. However, subtype B has also established successful epidemics in Asia and South America, and is the dominant type in Australia. Most likely has this subtype spread to these parts of the world from contacts with America and Europe. In Thailand a separate type of subtype B has been referred to as B' [26], and more recently B_B , which has been found to further spread into China, Malaysia, and Japan [10]. Subtype B and subtype E transmissions in Thailand have been compared, and some authors have proposed that in this community subtype E was found to be more transmissible [26, 32, 62]. This observation alarmed some European countries, where it was shown that subtype E had recently entered [2]. Subtype E has also entered the American continent, including the USA [5, 7], and subtype E variants have been found in Central Africa [44, 48].

Subtype C seems to have its major center in East and South Africa, but has been reported to occur in many other parts of the world, including Europe, Russia, China, India, and Brazil. As an effect of communication, geographical patterns are becoming more and more blurred for essentially all subtypes. This in turn will increase the possibility of superinfections of different subtypes, and thereby the virus will have greater opportunities to explore new genetic variants by recombination [46].

The O Group

Virus and sequences from the O group have been investigated less thoroughly, mainly because this type of virus is less common. Currently, 3 full length *env* gp120 sequences have been published and are available in the database, but several groups are investigating these viruses. Most of these viruses originate from individuals who have been infected in West Africa, or from encounters with someone infected in West Africa [27]. Analysis of data suggest that the O group sequences do not form subtype clusters in the same way as the M group sequences [39].

Methods of Investigation

Strategy. In any attempt to do a sequence study a good strategy is required. Depending on what the goal is, different approaches will apply. In many cases several areas of interest may have been pursued simultaneously, and in other cases the sequences may have been determined for a goal other than the present one. For instance, samples may have been collected from a restricted area or disease group, or may have gone through different in vitro manipulations; and sequences may have been derived from either proviral DNA or genomic RNA. All these factors should be considered before going further, since they may bias the conclusions.

Sequencing. After an investigation strategy has been decided and samples have been collected, a study often starts with determining the primary DNA sequence. As mentioned above, the sequence template may be either proviral DNA or reverse transcribed RNA. It is important to remember that any HIV sample contains a population of related but genetically diverse variants, or quasispecies [42] (see also [15, 16, 25]). Any sequencing strategy must take this fact into account, i.e. it must be certain that a representative number of viral molecules are investigated. If this is not done, the chance of making erroneous conclusions arises: for instance, one could detect a "homogeneous" population, or a special sequence, derived from a single odd molecule. Thus, the number of molecules that goes into the sequencing effort should be determined, for example by a limiting dilution procedure [60].

The easiest way to determine a virus sequence is by direct PCR sequencing. This procedure suffers the least from methodological artifacts by introducing the least amount of undesired selection steps. Because a population is sequenced, some nucleotide positions may be populated by more than

one state (a multistate position); for instance, a position can harbor both an A and a G. There are different ways to handle this information. Some authors simply choose the dominant form at each such site (e.g., [49]), while others choose to report both nucleotides to a certain threshold [37]. It has also been shown that the different nucleotides at such positions can be accurately quantified [37]. The main drawback of the direct population sequencing strategy is that the linkage between nucleotides in multistate positions will remain unknown in most cases.

The other way to determine a population is by cloning prior to sequencing. There are several ways to do this, but the two most common ways are vector cloning or limited dilution of PCR fragments. Vector cloning involves many steps in which the population structure may be skewed, including transformation, growth, and clone picking. The limited dilution procedure [60] avoids these steps, but instead requires many PCR reactions. Both procedures also suffer from the fact that PCR-induced mutations may be amplified and accidentally analyzed as true information. Another seldom addressed problem with clones is that they generate a much larger dataset. This may be cumbersome in the stage of phylogenetic inference, because the number of possible tree topologies quickly reaches astronomical numbers, thereby making more advanced methods out of the question. However, some studies require exact knowledge of individual molecules, and a cloning procedure becomes the method of choice.

In addition to sequencing, a number of other procedures exist to determine genotypes. These include restriction enzyme analysis, SSCP, and different probe protocols such as the heteroduplex mobility assay (HMA) [12]. The advantage of such methods is that they can screen large sample sets in short time; efforts have been made to produce reference sets for standardization of the HMA method [6]. The resolution of this method is, however, lower than that of a sequencing study. Methods for rapid screening of subtypes based on V3 peptide serotyping have also been developed [58].

An important factor to consider is the amount of information fed to the subsequent analysis, usually a tree reconstruction. Although the discussion of phylogenetic information is difficult, in principle, two components decide whether it will be a good reconstruction or not. The first is sequence length. It is evident that the ideal procedure would be to use the complete HIV genome. However, to sequence such a fragment is difficult, and therefore the sequences of subfragments are usually determined. The second component is information density. The env gene is known to display a much higher degree of variability than the *pol* gene. Thus, a 300 bp fragment covering a hypervariable region in *env* contains far more information than the same length of sequence in pol. Several other factors, such as convergent changes as a result of influence of the immune system and/or drug treatment will affect the quality of the information, but normally, if the length of the sequence is reasonably long, this will only yield a slight background noise. However, it may be wise to choose a region that does not contain extensive length variations, such as V1-V2 of *env*, because much information will be lost in alignment columns where these occur as a result of gapstripping procedures [33] that many phylogenetic inference methods include. Leitner et al. recently showed that a 285 bp fragment covering the *env* V3 region was sufficient to reconstruct a known intrasubtype transmission chain covering 13 years of viral evolution [35]. For classifying subtypes, less would be required because of the larger sequence differences. Figure 2 shows the genetic distance correlation of common subfragments to their complete gene sequences. As is evident from the figure, shorter sequences are essentially giving the same answer as longer ones. This was true for both intra and inter subtype relations. Interestingly, some of the subfragments carry a higher information density than the complete genes: the V3 fragment of *env* and the p17 fragment of *gag* both display longer genetic distances than the complete genes (Figures 2B and 2C). Note, however, that the shorter a fragment gets, the more it will be affected by mutational noise and the lower the significance will be. Yet another problem with too-short sequences is that there may not be enough information there to estimate some parameters for more advanced distance estimates.

Phylogenetic inference. When a DNA sequence has been determined for a sample, the next step is to align it to the set of reference sequences describing established subtypes. Alignments of sequences covering the complete *env*, *gag*, and *pol* genes as well as subfragments gp120 and V3 of *env*, p17 (MA) of *gag*, and p51 (RT) of *pol* of the suggested reference sequences (Table 1) are available at the HIV database (http://hiv-web.lanl.gov). For the complete gene sequences, sets containing the subtypes shown in Figure 3 are included, while for the subfragments also more uncommon subtypes are represented when such sequences were available (according to Table 1).

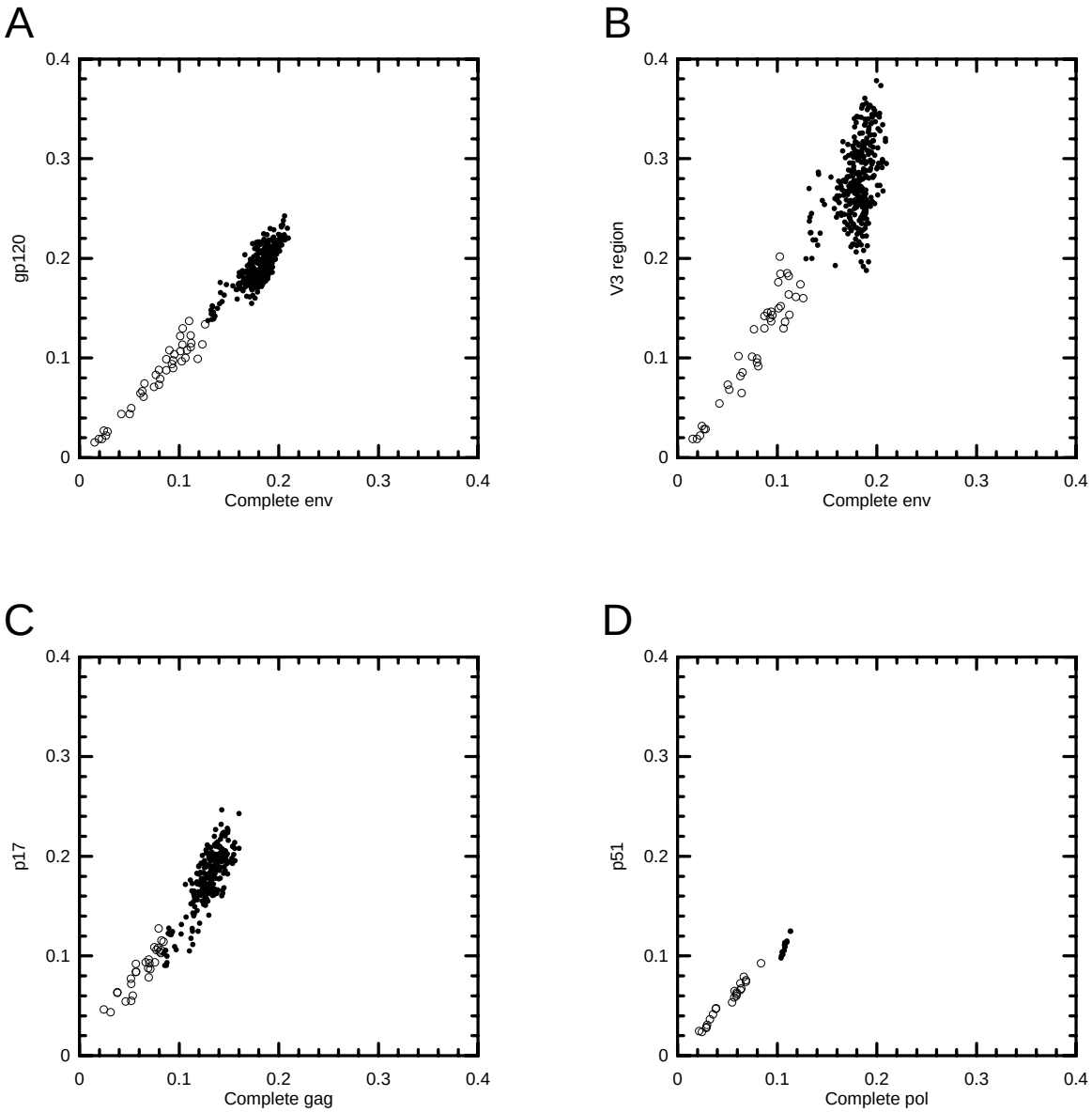

Figure 2. Comparison of distance estimates of nucleotide sequences from complete gene fragments to common gene subfragments. Open circles indicate within-subtype distances, and solid circles between-subtype distances. (A) The gp120 coding region (1251 bp) compared to the complete *env* gene (2301 bp); (B) The V3 region of gp120 (324 bp) compared to the complete *env* gene; (C) The p17 fragment of *gag* (381 bp including 24 bp into p24) compared to the complete *gag* gene (1380 bp); (D) The p51 fragment of *pol* (1320 bp) compared to the complete pol gene (3000 bp). All distance estimates were calculated using the F84 ($\kappa = 2.0$) nucleotide substitution model [19]. Note that sequence lengths are given for globally gapstripped datasets.

Figure 3 shows how maximum-likelihood, maximum-parsimony, and minimum-evolution based phylogenetic methods perform in reconstructing the established subtypes by use of *env*, *gag*, or *pol* sequences. All these methods and fragments can reconstruct the established subtypes. These trees were all calculated using the full length gene sequences, but essentially the same result would have been

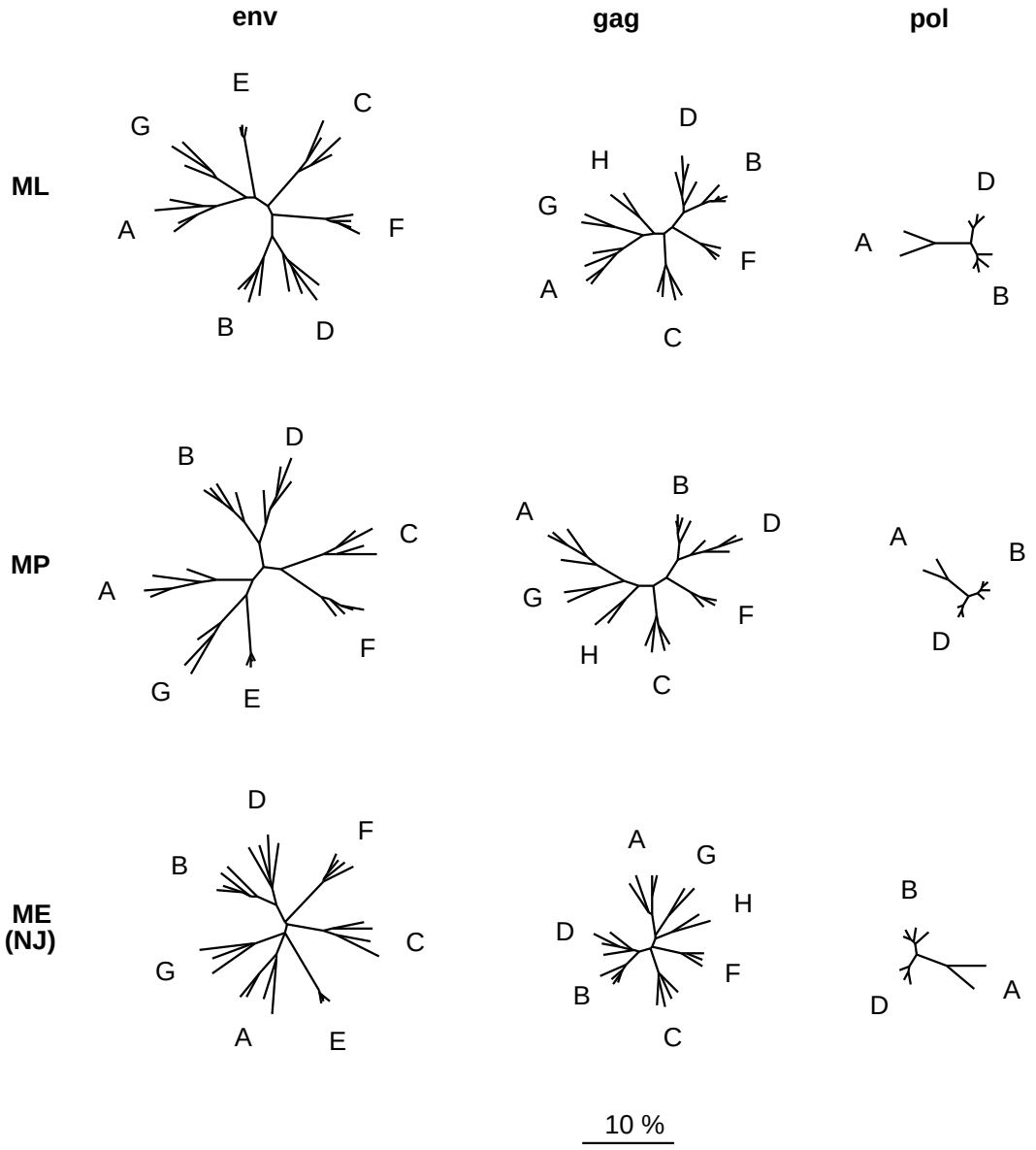

Figure 3. Unrooted trees calculated according to the maximum likelihood (ML) criteria, the maximum parsimony (MP) criteria, and the minimum evolution (ME) criteria. The ML search was conducted using the PHYLIP package [19] with the F84 substitution model, the MP search was conducted with a heuristic search using the PAUP program [63] under uniform weighting, and the ME search was conducted with the neighbor joining approximation using the PHYLIP package [19] with the F84 substitution model. Each method was applied to the complete *env*, *gag*, and *pol* gene alignments, gapstripped to 2301, 1380, and 3000 nucleotide sites, respectively. Branch lengths in ML and ME trees were adjusted to the indicated scale bar, and MP trees were drawn arbitrarily to reflect the information content in a comparable scale.

derived using the suggested subfragments (Figure 2). Thus, for subtyping purposes methods based on these criteria (maximum-likelihood, maximum-parsimony, and minimum-evolution) will give an acceptable answer. In other situations the method may influence the result. However, it has recently been shown in detailed analyses of transmission cases that the amount of information is more important

than the choice of phylogenetic inference method [35]. The opposite is true for calculating genetic distances [38]. Although this is not a primary goal when attempting to determine the subtype of a sequence, it may provide useful information regarding relatedness. When a p-distance (proportion of differences) reaches above 0.1, the choice of substitution model is already making a major impact on the distance estimates. From here, the effect of superimposed events is then rapidly increasing by increased distance. By using an oversimplified model of nucleotide substitution, longer distances will be less and less accurate. With increased distance, saturation effects also start to appear, especially on synonymous positions. If the sequence is long enough, one may consider using only nonsynonymous positions, only second codon positions, or translated amino acid sequences. Again, however, for simple subtyping purposes the substitution model is unlikely to influence the results.

Figure 3 also indicates the information density of the three investigated genes. As is evident from all trees, independent of phylogenetic method, the branch lengths of the tree are not reflected by the sequence length, but rather by the density of nucleotide variation of each gene. In other words, the faster a gene evolves the longer the branch lengths will be, as compared to the more slowly evolving gene over the same time interval. The speed at which the genes are evolving is dictated by several factors, most notably mutation and selection.

Rooting or using a specific sequence as an outgroup in a tree analysis, makes no difference when determining the subtype of a sequence. Depending on which type of tree plot one chooses to present the results, however, the choice of outgroup may influence the aesthetics of the tree. A root should only be assessed when its location in the phylogeny is known, while an outgroup can be chosen arbitrarily [64].

When a sequence falls outside an established subtype, more analyses are needed. The first thing to do would be to check if the used sequence fragment includes any sign of recombination. This can be done by several methods, including signature analysis by VESPA [17, 28] or RIP [59], tree analysis by the bootscanning method [56], or even by simply eyeballing an alignment against representative subtype sequences or subtype consensus sequences. A more extensive tree analysis should also be done, including all available sequences from the two subtypes surrounding the odd sequence as well as earlier reported unclassified sequences. The next step would be to sequence another part of the HIV genome, to see if the classification there indicates an odd location as well. Congruence of trees from two separate coding regions is a strong indication of a new genotype. Also, one should check that the sequence analyzed is not odd because of hypermutation events [66]. If the sequence falls outside established subtypes in two regions, and no signs of recombination are seen, it would be desirable to sequence the complete genome to allow for extensive recombination analysis and/or to establish a new subtype (or potential subtype, if only one sample was found).

Bootstrapping is a resampling technique used to test the robustness of an inferred tree topology [13, 18]. The method resamples data points from the original dataset, and for each resampled dataset a tree is inferred. Finally a consensus tree is constructed and each cluster in this tree is assigned a “bootstrap value” corresponding to the number of times it occurred in the resampled trees. The interpretation of such values varies among authors, since the values provide unbiased but highly imprecise estimates of repeatability. Some authors have claimed that the probability that the result represents the true phylogeny is a biased, but usually conservative, estimate of accuracy [24]. Recently, however, other authors showed that the method is not biased, but that it can be corrected to better agree with standard ideas of confidence levels and hypothesis testing [14]. The degree of “bias” varies from node to node, as well as from study to study, so bootstrap values are not directly comparable to each other. Additionally, bootstrap values for a given cluster in any tree will be highly affected by the control sequences chosen. However, many subtype studies include this analysis, and at best it gives a measure of how stable the inference is. In general, all the established subtypes should present high bootstrap values, but congruence of trees from two coding regions are preferable to bootstrap analysis over just one region.

Implications and Epidemiology

Genetic subtypes were originally defined on *env* and *gag* sequences. The main goal was to identify the major genetic lineages [40, 45]. Since then many studies have included or even concentrated on efforts of subtype classifications. The presence of distant genotypes also made it possible to investigate and realize that recombination events (and thereby, superinfections) are common, or at least not uncommon, in HIV evolution [36, 50, 52]. However, the key question that has been raised among researchers in this field is: What does a genetic subtype mean?

There are three main reasons for why genetic subtypes have been considered to be important to investigate, where the first two are based on the belief that the viral genetic factor is important in virus to host interactions. First, in the context of vaccine development it was realized that the different subtypes express different envelope proteins, and that this likely would effect vaccine treatments. For instance, the tip of the principal neutralization domain of the V3 loop is GPGR in subtype B viruses, while in most other subtypes the consensus is GPGQ. Together with the knowledge that genetic variants can escape the immune system, information on genetic subtypes becomes important [8, 26]. Another similar problem, resulting from genetic variation among subtypes is that some methods can fail in detecting virus [3].

Second, several studies have tried to correlate genetic factors to biological properties. For instance, specific amino acids in the *env* V3 loop have been found to be markers for the viral phenotype [11, 20]. However, such a correlation have been found to be subtype independent, or in other words, different phenotypes exist in all genetic subtypes [36, 51, 68]. Kostrikis et al. recently showed that no direct correlation between neutralization serotypes and genetic subtypes could be established [31]. It has been suggested that the Thai subtype E may be more transmissible than subtype B in Thailand [32, 62], but it is difficult to discriminate between the viral genetic factor and the epidemiological situation in such a comparison. Recently, viral phenotype was found to be dominant over genotype also in the context of chemokine receptors as being the long-sought second receptor of HIV-1 [67]. Genotypic studies have often concentrated to regions of special interest, and it is therefore not clear if genetic factors elsewhere in the HIV genome may play an important biological role. However, mutational trends have been observed within the V3 loop among different subtypes, which may reflect the acquisition of specific biological properties during the evolution of HIV-1 [29].

Finally, more informative are the subtypes for the understanding of the HIV-1 epidemic. It was early realized that subtype B represented the vast majority of virus circulating in North America and Europe, while it was almost absent in Africa. Instead, all other subtypes were found in Africa (more recently, subtype B has been reported several times from different African countries). Together with the fact that many species of African monkeys carry similar viruses, it was concluded that the origin of HIV was somewhere in Africa. The question of when the cross-species transmission from monkeys to humans occurred has been investigated by several groups, and the answers have been astonishingly different: ranging from 40 to millions of years ago [22, 57, 61]. The main reason for the different interpretations is that the true evolutionary pattern of the virus is not known, creating problems of interpreting the shape of the tree and correlating time to branch lengths. However, the current subtypes do tell us that very different viral genotypes can rapidly establish successful epidemics in naive human populations. This suggests that each subtype is well-adapted to its environment, but also that any existing variant could have established a successful clade. In this context, it is interesting to look a bit closer to the actual shape of the subtype tree (Figures 1 and 3). We can see that the star-like shape actually has structure: subtype B and D are always closer to each other; subtype A is usually more divergent than other subtypes and normally close to G and H; and that more recent subtypes form more distinct clusters away from clusters that mainly are derived from Central African sequences. It seems like there are two components in the tree: an early phase in which the virus spread more slowly between host groups in Africa, forming its network of variants; and a recent phase, which we are still in, in which the virus reaches a naive human population and uses its standing network to spread rapidly. Perhaps can this teach us about epidemic spreads, and how to prevent them in the future.

References

[1] Alaeus, A., T. Leitner, K. Lidman, and J. Albert. 1996. Most genetic subtypes of HIV-1 have entered Sweden. *AIDS*:in press.

[2] Anonymous. 1996. Also HIV subtype E is not transmitted by kissing. *Z. Arztl. Fortbild. (Jena)*. **90**:49 (in German).

[3] Arnold, C., K. L. Barlow, S. Kaye, C. Loveday, P. Balfe, and J. P. Clewley. 1995. HIV type 1 sequence subtype G transmission from mother to infant: failure of variant sequence species to amplify in the Roche Amplicor Test. *AIDS Res. Hum. Retrovirus*. **11**:999–1001.

[4] Arnold, C., K. L. Barlow, J. V. Parry, and J. P. Clewley. 1995. At least five HIV-1 subtypes (A, B, C, D, A/E) occur in England. *AIDS Res. Hum. Retrovirus*. **11**:427–429.

[5] Artenstein, A. W., C. A. Ohl, T. C. VanCott, P. A. Hegerich, and J. R. Mascola. 1996. Transmission of HIV-1 subtype E in the united states. *JAMA*. **276**:99–100.

[6] Bachmann, M. H., E. L. Delwart, E. G. Shpaer, P. Lingenfelter, R. Singal, and J. I. Mullins. 1994. Rapid genetic characterization of HIV type 1 strains from four World Health Organization-sponsored vaccine evaluation sites using a heteroduplex mobility assay. WHO Network for HIV Isolation and Characterization. *AIDS Res. Hum. Retrovirus*. **10**:1345–1353.

[7] Brodine, S. K., J. R. Mascola, P. J. Weiss, S. I. Ito, K. R. Porter, A. W. Artenstein, F. C. Garland, F. E. McCutchan, and D. S. Burke. 1995. Detection of diverse HIV-1 genetic subtypes in the USA. *Lancet*. **346**:1198–1199.

[8] Buonaguro, L., E. Del Guadio, M. Monaco, D. Greco, P. Corti, E. Beth-Giraldo, F. M. Buonaguro, and G. Giraldo. 1995. Heteroduplex mobility assay and phylogenetic analysis of V3 region sequences of human immunodeficiency virus type 1 isolates from Gulu, northern Uganda. The Italian-Ugandan Cooperation AIDS Program. *J. Virol*. **69**:7971–7981.

[9] Carr, J. K., M. O. Salminen, C. Koch, D. Gotte, A. W. Artenstein, P. A. Hegerich, D. St. Louis, D. S. Burke, and F. E. McCutchan. 1996. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J. Virol*. **70**:5935–5943.

[10] Cassol, S., B. G. Weniger, P. G. Babu, M. O. Salminen, X. Zheng, M. T. Htoon, A. Delaney, M. O'Shaughnessy, and C.-Y. Ou. 1996. Detection of HIV type 1 env subtypes A, B, C, and E in Asia using dried blood spots: a new surveillance tool for molecular epidemiology. *AIDS Res. Hum. Retrovirus*. **12**:1435–1441.

[11] De Jong, J. J., A. De Ronde, W. Keulen, M. Tersmette, and J. Goudsmit. 1992. Minimal requirements for the human immunodeficiency virus type 1 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J. Virol*. **66**:6777–6780.

[12] Delwart, E. L., E. G. Shpaer, J. Louwagie, F. E. McCutchan, M. Grez, H. Rübsamen-Waigmann, and J. I. Mullins. 1993. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science*. **262**:1257–1261.

[13] Efron, B. 1982. The jackknife, the bootstrap, and other resampling techniques. CBMS-NFS Regional Conference Series in Applied Mathematics, Monograph 38. Society of Industrial and Applied Mathematics, Philadelphia.

[14] Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*. **93**:13429–13434.

[15] Eigen, M., J. McCaskill, and P. Schuster. 1988. Molecular quasi-species. *J. Phys. Chemis*. **92**:6881–6891.

[16] Eigen, M., and R. Winkler-Oswatitsch. 1990. Statistical geometry in sequence space. *Meth. Enymol*. **183**:505–530.

[17] Farmer, A., and G. Myers. 1995. Detection of HIV hybrid sequences using VESPA. In G. Myers and B. Korber and B. H. Hahn and K.-T. Jeang and J. W. Mellors and F. E. McCutchan and L. E. Henderson and G. N. Pavlakis (ed.), *Human Retroviruses and AIDS 1995: A compilation of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[18] Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. **39**:783–791.

[19] Felsenstein, J. 1993. PHYLIP: Phylogeny Inference Package, 3.52c ed. University of Washington, Seattle, WA.

[20] Fouchier, R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable region of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol*. **66**:3183–3187.

[21] Fransen, K., A. Buve, J. N. Nkengasong, M. Laga, and G. van der Groen. 1996. Longstanding presence in Belgians of multiple non-B HIV-1 subtypes. *Lancet*. **347**:1403.

[22] Fukasawa, M., T. Miura, A. Hasegawa, S. Morikawa, H. Tsujimoto, K. Miki, T. Kitamura, and M. Hayami. 1988. Sequence of simian immunodeficiency virus from African green monkey, a new member of the HIV/SIV group. *Nature*. **333**:457–461.

[23] Gao, F., S. G. Morrison, D. L. Robertson, C. L. Thornton, S. Craig, G. Karlsson, J. Sodroski, M. Morgado, B. Galvao-Castro, H. von Briesen, S. Beddows, J. Weber, P. M. Sharp, G. M. Shaw, and B. H. Hahn. 1996. Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. The WHO and NIAID Networks for HIV Isolation and Characterization. *J. Virol*. **70**:1651–1657.

[24] Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol*. **42**:182–192.

[25] Holland, J. J., J. C. De La Torre, and D. A. Steinhauer. 1992. RNA virus populations as quasispecies. *Curr. Top. Microbiol. Immunol*. **176**:1–20.

[26] Kalish, M. L., A. Baldwin, S. Raktham, C. Wasi, C.-C. Luo, G. Schochetman, T. D. Mastro, N. L. Young, S. Vanichseni, H. Rubsamen-Waigmann , H. von Briesen, J. I. Mullins, E. Delwart, B. Herring, J. Esparza, W. L. Heyward, and S. Osmanov. 1995. The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. *AIDS*. **9**:851–857.

[27] Korber, B., I. Loussert-Ajakai, J. Blouin, and S. Saragosti. 1996. A comparison HIV-1 M and group O functional and immunogenic domains in the gag p24 protein and the C2V3 region of the envelope protein, p. III-39. Published in this compendium.

[28] Korber, B., and G. Myers. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res. Hum. Retrovirus*. **8**:1549–1558.

[29] Korber, B. T. M., K. MacInnes, R. F. Smith, and G. Myers. 1994. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. *J. Virol*. **68**:6730–6744.

[30] Kostrikis, L. G., E. Bagdades, Y. Cao, L. Zhang, D. Dimitriou, and D. D. Ho. 1995. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. *J. Virol*. **69**:6122–6130.

[31] Kostrikis, L. G., Y. Cao, H. Hgai, J. P. Moore, and D. D. Ho. 1996. Quantitative analysis of serum neutralization of human immunodeficiency virus type 1 from subtypes A, B, C, D, E, F, and I: lack of direct correlation between neutralization serotypes and genetic subtypes and evidence for prevalent serum-dependent infectivity enhancement. *J. Virol*. **70**:445–458.

[32] Kunanusont, C., H. M. Foy, J. K. Kreiss, S. Rerks-Ngarm, P. Phanuphak, S. Raktham, C.-P. Pau, and N. L. Young. 1995. HIV-1 subtypes and male-to-female transmission in Thailand. *Lancet*. **345**:1078–1083.

[33] Learn, G. H. J., B. T. M. Korber, B. Foley, B. H. Hahn, S. M. Wolinsky, and J. I. Mullins. 1996. Maintaining the integrity of human immunodeficiency virus sequence databases. *J. Virol*. **70**:5720–5730.

[34] Leitner, T. 1996. Genetic variation of HIV-1: Molecular epidemiology and viral evolution. Swedish Institute for Infectious Disease Control, Karolinska Institute, Stockholm, Sweden.

[35] Leitner, T., D. Ecanilla, C. Franzén, M. Uhlén, and J. Albert. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. USA*. **93**:10864–10869.

[36] Leitner, T., D. Escanilla, S. Marquina, J. Wahlberg, C. Broström, H. B. Hansson, M. Uhlén, and J. Albert. 1995. Biological and molecular characterization of subtype D, G and A/D recombinant HIV-1 transmissions in Sweden. *Virology*. **209**:136–146.

[37] Leitner, T., E. Halapi, G. Scarlatti, P. Rossi, J. Albert, E. M. Fenyö, and M. Uhlén. 1993. Analysis of heterogeneous viral populations by direct DNA sequencing. *BioTechniques*. **15**:120–126.

[38] Leitner, T., S. Kumar, and J. Albert. 1996. Tempo and mode of the nucleotide substitution in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. Submitted.

[39] Loussert-Ajakai, I., M.-L. Chaix, B. Korber, F. Letourneur, E. Gomas, T.-D. Ly, F. Brun-Vezinet, F. Simon, and S. Saragosti. 1995. The variability of HIV type 1 group O strains isolated from Cameroonian patients living in France. *J. Virol*. **69**:5640–5649.

[40] Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, B. E. Sanders, G. A. Eddy, G. van der Groen, K. Fransen, G.-M. Gershy-Damet, R. Deleys, and D. S. Burke. 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS*. **7**:769–780.

[41] Lukashov, V. V., M. T. Cornelissen, J. Goudsmit, M. N. Papuashvilli, P. G. Rytik, R. M. Khaitov, E. V. Karamov, and F. de Wolf. 1995. Simultaneous introduction of distinct HIV-1 subtypes into risk groups in Russia, Byelorussia and Lithuania. *AIDS*. **9**:435–439.

[42] Meyerhans, A., R. Cheynier, J. Albert, M. Seth, S. Kwok, J. Sninsky, L. Morfeldt-Månson, B. Åsjö, and S. Wain-Hobson. 1989. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell*. **58**:901–910.

[43] Montpetit, M. 1995. HIV-1 subtype A in Canada. *AIDS Res. Hum. Retrovirus*. **11**:1421–1422.

[44] Murphy, E., B. Korber, M. C. Georges-Courbot, B. You, A. Pinter, D. Cook, M. P. Kieny, A. Georges, C. Mathiot, F. Barre-Sinoussi, and M. Girard. 1993. Diversity of V3 region sequences of Human Immunodeficiency viruses type-1 from the Central-African-Republic. *AIDS Res. Hum. Retrovirus*. **9**:997–1006.

[45] Myers, G. 1993. Assimilating HIV sequences. *AIDS Res. Hum. Retrovirus*. **9**:697–702.

[46] Myers, G., B. Korber, B. H. Hahn, K.-T. Jeang, J. W. Mellors, F. E. McCutchan, L. E. Henderson, and G. N. Pavlakis. 1995. *Human retroviruses and AIDS 1995: a compilation and analysis of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[47] Myers, G., B. Korber, S. Wain-Hobson, K.-T. Jeang, L. E. Henderson, and G. N. Pavlakis. 1994. *Human retroviruses and AIDS 1994: a compilation and analysis of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[48] Nkengasong, J. N., W. Janssens, L. Heyndrickx, K. Fransen, P. M. Ndumbe, J. Motte, A. Leonaers, M. Ngolle, J. Ayuk, P. Piot, and G. van der Groen. 1994. Genotypic subtypes of HIV-1 in Cameroon. *AIDS*. **8**:1405–1412.

[49] Ou, C. Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C.-C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. MacInnes, J. W. Curran, H. W. Jaffe, Laboratory investigation group, and Epidemiologic investigation group. 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science*. **256**:1165–1171.

[50] Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn. 1995. Recombination in HIV-1. *Nature*. **374**:124–126.

[51] Rübsamen-Waigmann, H., H. von Briesen, H. Holmes, A. Björndal, B. Korber, R. Esser, S. Ranjbar, P. Tomlinson, B. Galvao-Castro, E. Karita, S. Sempala, C. Wasi, S. Osmanov, and E. M. Fenyö. 1994. Standard conditions of virus isolation reveal biological variability of HIV type 1 in different regions of the world. WHO Network for HIV Isolation and Characterization. *AIDS Res. Hum. Retrovirus*. **10**:1401–1408.

[52] Sabino, E. C., E. G. Shpaer, M. G. Morgado, B. T. M. Korber, R. S. Diaz, V. Bongertz, S. Cavalcante, B. Galvão-Castro, J. I. Mullins, and A. Mayer. 1994. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J. Virol*. **68**:6340–6346.

[53] Salminen, M. 1996. Personal communication.

[54] Salminen, M., A. Nykänen, H. Brummer-Korvenkontio, M. L. Kantanen, K. Liitsola, and P. Leinikki. 1993. Molecular epidemiology of HIV-1 based on phylogenetic analysis of in vivo gag p7/p9 direct sequences. *Virology*. **195**:185–194.

[55] Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan. 1995. Genotyping of HIV-1, p. III30–III34. In G. Myers and B. Korber and B. H. Hahn and K.-T. Jeang and J. W. Mellors and F. E. McCutchan and L. E. Henderson and G. N. Pavlakis (ed.), *Human retroviruses and AIDS 1995: a compilation and analysis of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM.

[56] Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan. 1995. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retrovirus*. **11**:1423–1425.

[57] Sharp, P. M., D. L. Robertson, F. Gao, and B. H. Hahn. 1994. Origins and diversity of human immunodeficiency viruses. *AIDS*. **8** (suppl 1):S27–S42.

[58] Sherefa, K., A. Sonnerborg, J. Steinbergs, and M. Sallberg. 1994. Rapid grouping of HIV-1 infection in subtypes A to E by V3 peptide serotyping and its relation to sequence analysis. *Biochem. Biophys. Res. Commun*. **205**:1658–1664.

[59] Siepel, A. C., A. L. Halpern, C. Macken, and B. Korber. 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retrovirus*. **11**:1413–1416.

[60] Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop, and A. J. Leigh Brown. 1990. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. *J. Virol*. **64**:864–872.

[61] Smith, T. F., A. Srinivasan, G. Schochetman, M. Marcus, and G. Myers. 1988. The phylogenetic history of immunodeficiency viruses. *Nature*. **333**:573–575.

[62] Soto-Ramirez, L. E., B. Renjifo, M. F. McLane, R. Marlik, C. O'Hara, R. Sutthent, C. Wasi, P. Vithayasai, V. Vithayasai, C. Apichartpiyakul, P. Auewarakul, V. Peña Cruz, D.-S. Chui, R. Osathanondh, K. Mayer, T.-H. Lee, and M. Essex. 1996. HIV-1 Langerhans' cell tropism associated with heterosexual transmission of HIV. *Science*. **271**:1291–1293.

[63] Swofford, D. L. 1991. PAUP: Phylogenetic Analysis Using Parsimony, 3.1.1 ed. Illinois Natural History Survey, Champaign, Illinois.

[64] Swofford, D. L., and G. J. Olsen. 1990. Phylogeny reconstruction, p. 411–501. In D. M. Hillis and C. Moritz (ed.), *Molecular systematics*. Sinauer Associates, Sunderland, MA.

[65] Tsuchie, H., T. S. Saraswathy, M. Sinniah, B. Vijayamalar, J. K. Maniar, O. T. Monzon, R. T. Santana, F. J. E. Paladin, C. Wasi, P. Thongcharoen, M. M. Hossain, O. Yamada, S. Kageyama, T. Kitamura, and T. Kurimura. 1995. HIV-1 variants in South and South-East Asia. *Int. J. STD AIDS*. **6**:117–120.

[66] Vartanian, J.-P., A. Meyerhans, M. Sala, and S. Wain-Hobson. 1994. G - A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. *Proc. Natl. Acad. Sci. USA*. **91**:3092–3096.

[67] Zhang, L., Y. Huang, T. He, Y. Cao, and D. D. Ho. 1996. HIV-1 subtype and second-receptor use. *Nature*. **383**:768.

[68] Zhong, P., M. Peeters, W. Janssens, K. Fransen, L. Heyndrickx, G. Vanham, B. Willems, P. Piot, and G. van der Groen. 1995. Correlation between genetic and biological properties of biologically cloned HIV type 1 viruses representing subtypes A, B, and D. *AIDS Res. Hum. Retrovirus*. **11**:239–248.